Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 04

Pierre-Luc Germain

ETH Zürich

Plan for today

- Debriefing on the assignment
- Recap of last week
- Manipulating and visualizing peaks
- Coverage track generation
- ENCODE & functional elements
- Finding data from the literature

Debriefing on the assignments

- Handing in the exercises etc.:
  - Handing in the exercises: Please name the exercises files just **`assignment.html`**
  - *From now on points will be deduced if the file is not correctly named*

- When re-using code from the practicals, you should update the filenames :

```
download.file("https://www.encodeproject.org/files/ENCFF127RRR/@@download/ENCFF127RRR.fastq.gz", dest="raw/Myc.fastq.gz", mode = "wb")
```

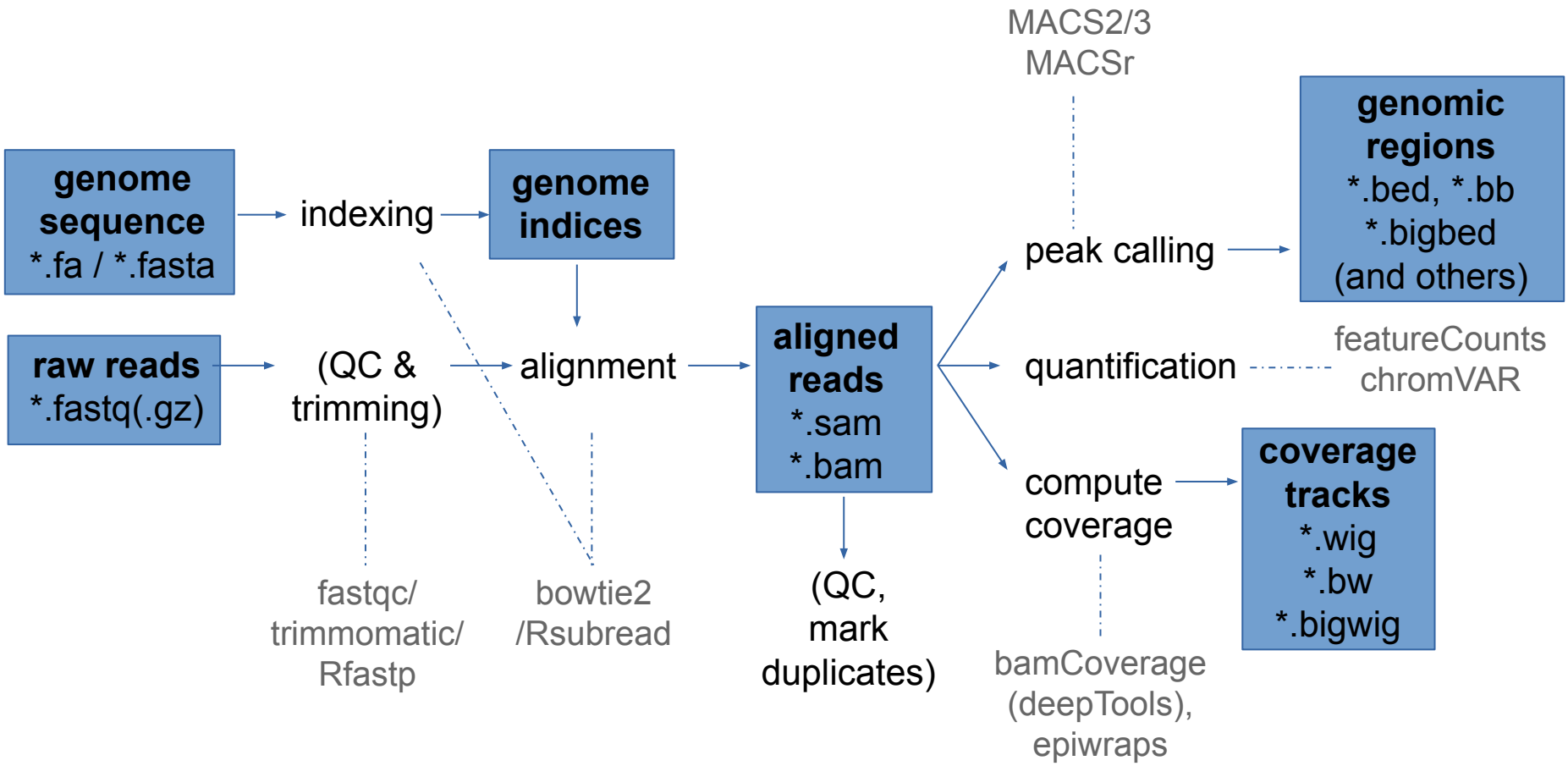
```
align.stats <- Rsubread::align(index="BDGP6_genome/rsubread", type="dna",
                               readfile1="rfastp.trimmed/Myc_R1.fastq.gz",
                               output_file="aligned/Myc.bam",
                               nthreads=6, sortReadsByCoordinates=TRUE)
```

=> however what we downloaded for this exercise is a **CTCF** dataset

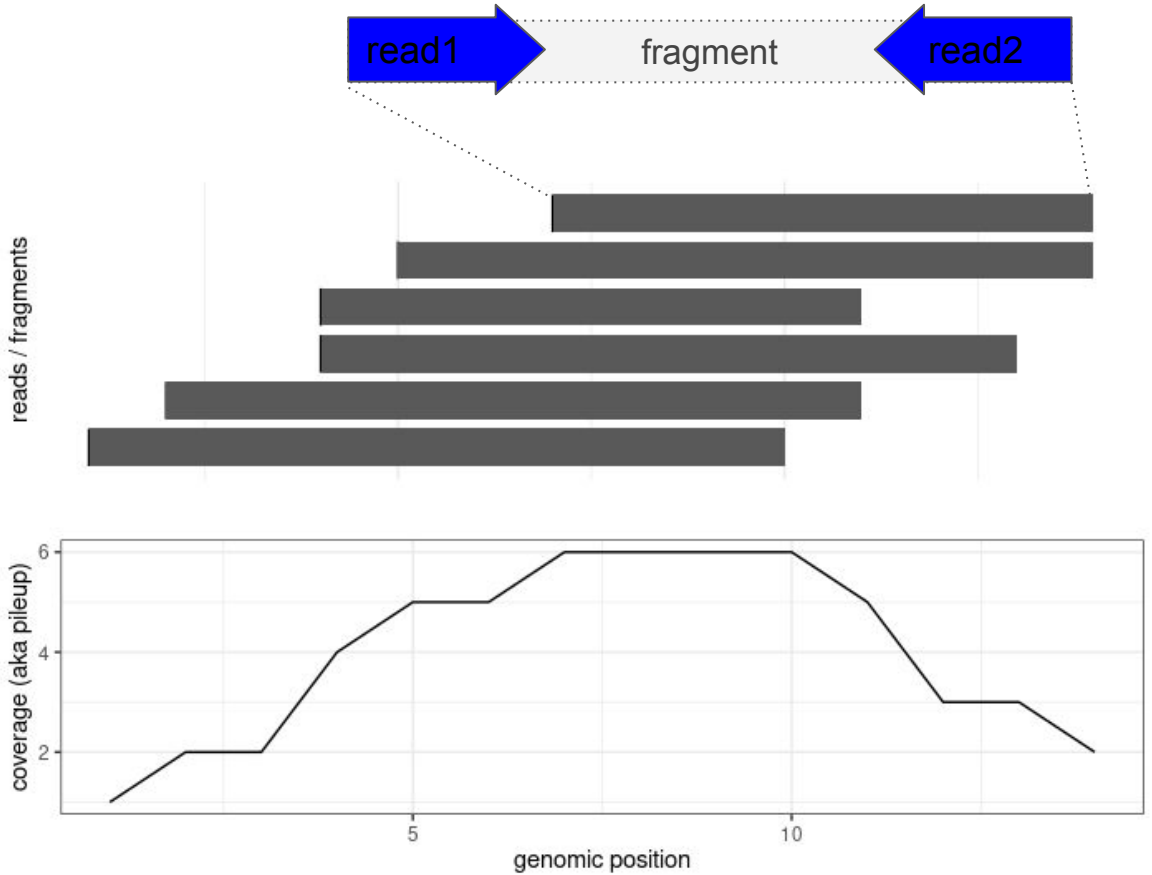
Can easily lead to mistakes if you have several files on disk.

```
download.file("https://www.encodeproject.org/files/ENCFF127RRR/@@download/ENCFF127RRR.fastq.gz", dest="raw/CTCF.fastq.gz")
```

Overview of a primary analysis pipeline (ChIP-seq and the likes)

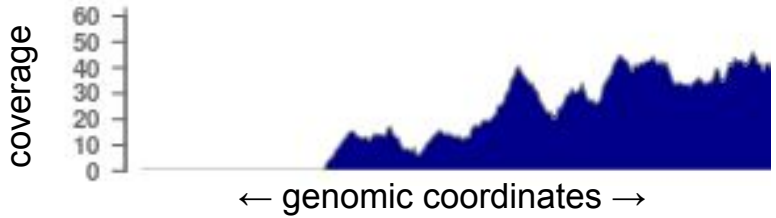


Recap of fragment summarization



Visualizations available in *epiwraps*

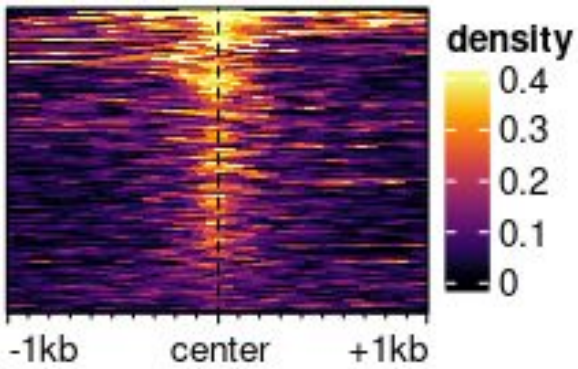
- Signal across one genomic region: `plotSignalTracks`



- `Input: bam/bigwig/bed/GRanges`

(Based on the *Gviz* R package)

- Signal across several genomic regions: `signal2Matrix` → `plotEnrichedHeatmaps`

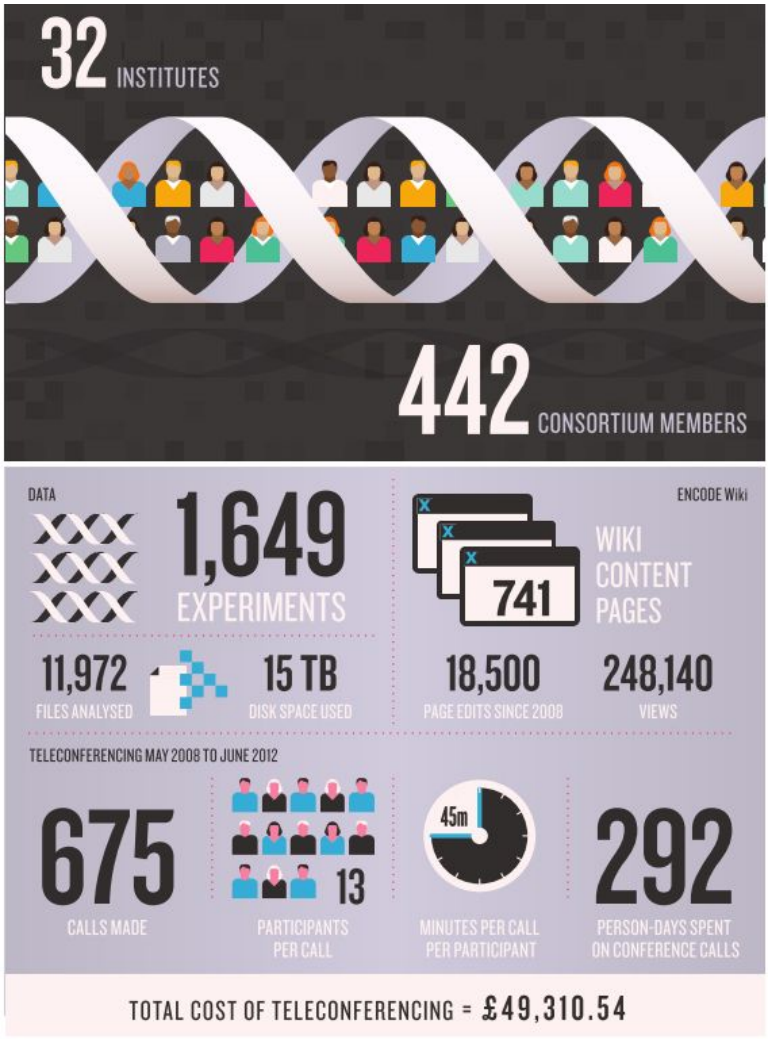


(Mainly based on the EnrichedHeatmap R package, itself based on ComplexHeatmap)

Read extension in coverage track generation

Coverage without read extension

Coverage with read extension



The ENCyclopedia Of DNA Elements

~30 publications in September 2012

$288 million USD

… then an ENCODE2, 3, now working towards the 5…

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium

Nature **489**, 57–74 (2012) | Cite this article

Bits of Mystery DNA, Far From 'Junk,' Play Crucial Role

The New York Times

by Gina Kolata

"At least 80 percent of this DNA is *active* and *needed*."

The evolutionary arguments for junk:

- 1% protein-coding
- ~4 to 10% evolutionarily conserved
- >50% transposable elements
- Onions have a 5 times bigger genome

The very angry response:

- Graur et al., GBE 2013

NEWS&ANALYSIS

GENOMICS

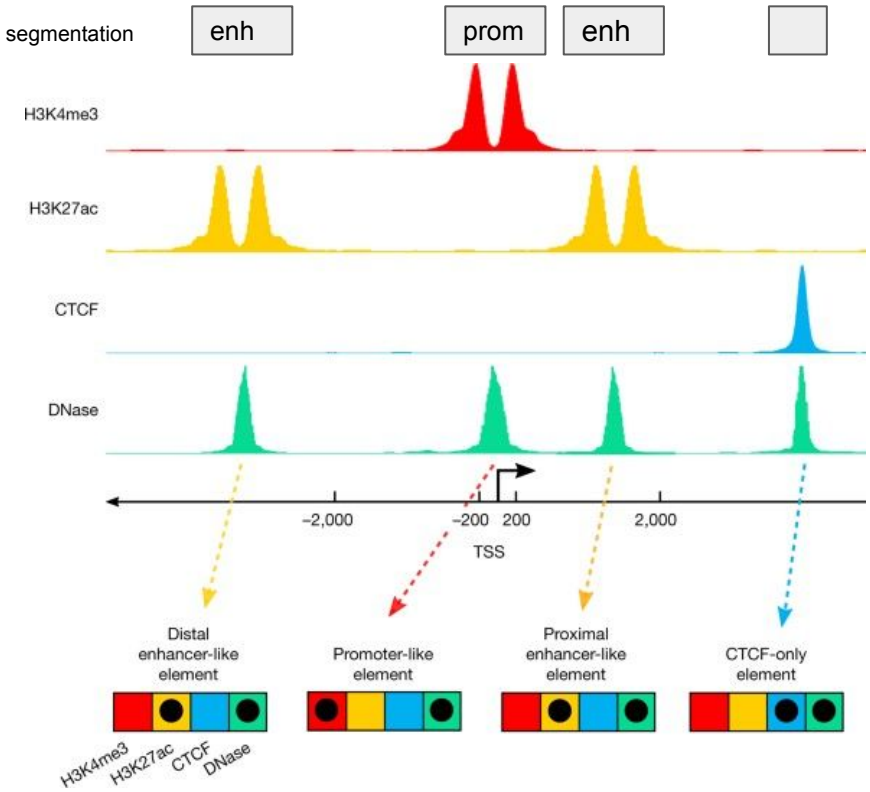
ENCODE Project Writes Eulogy For Junk DNA

–ELIZABETH PENNISI

SCIENCE VOL 337 7 SEPTEMBER 2012

(For more on the topic, see Germain, Ratti and Boem 2014; Ratti and Germain 2022)

A signature-based encyclopedia of DNA elements



ENCODE's "signature strategy":

- Different types of functional genetic elements are associated with different chemical signatures
- We can identify functional elements by identifying these signatures genome-wide

Generic repositories for NGS data

https://www.ebi.ac.uk/biostudies/arrayexpress

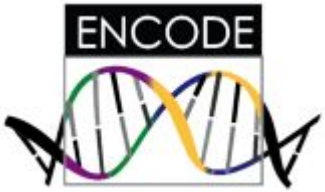
https://www.ncbi.nlm.nih.gov/geo/

https://www.ebi.ac.uk/ena/

https://www.ncbi.nlm.nih.gov/sra

International Nucleotide Sequence Database Collaboration

Quality-controlled and uniformly processed human and mouse NGS datasets



www.roadmapepigenomics.org

www.encodeproject.org

(hematopoietic system)

Assignment

- Find and download from ENCODE the **peaks** (i.e. bed-like format) for the following histone modifications in mouse embryonic stem cells (mESC) from ENCODE:
  - p300, H3K4me3, H3K4me1, H3K27ac, and H3K27me3
  - (when there are replicates, we recommend using the bed file denoted as "`conservative IDR thresholded peaks`")
- Of the p300 peaks, what proportion overlap each of the marks?
- Don't forget to upload your assignment as "`assignment.html`" !